The Australian National University Semester 2, 2021
School of Computing Theory Assignment 2 of 4
Liang Zheng

COMP3670/6670: Introduction to Machine Learning

Release Date. 18th August 2021
Due Date. 23:59pm, 19th September 2021
Maximum credit. 100

Errata: In Exercise 4, the loss function included a regulariser term $\|\mathbf{c}\|_{\mathbf{B}}^2$, which is undefined due to a dimensionality mismatch. This has been replaced with $\|\mathbf{c}\|_{\mathbf{A}}^2$.

Exercise 1 **Inner Products induce Norms** 20 credits

Let V be a vector space, and let $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ be an inner product on V. Define $||\mathbf{x}|| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. Prove that $|| \cdot ||$ is a norm.

(Hint: To prove the triangle inequality holds, you may need the Cauchy-Schwartz inequality, $\langle \mathbf{x}, \mathbf{y} \rangle \leq ||\mathbf{x}||||\mathbf{y}||$.)

Exercise 2 **Vector Calculus Identities** 10+10 credits

1. Let $\mathbf{x}, \mathbf{a}, \mathbf{b} \in \mathbb{R}^n$. Prove that $\nabla_{\mathbf{x}}(\mathbf{x}^T \mathbf{a} \mathbf{b}^T \mathbf{x}) = \mathbf{a}^T \mathbf{x} \mathbf{b}^T + \mathbf{b}^T \mathbf{x} \mathbf{a}^T$.
2. Let $\mathbf{B} \in \mathbb{R}^{n \times n}, \mathbf{x} \in \mathbb{R}^n$. Prove that $\nabla_{\mathbf{x}}(\mathbf{x}^T \mathbf{B} \mathbf{x}) = \mathbf{x}^T(\mathbf{B} + \mathbf{B}^T)$.

Exercise 3 **Properties of Symmetric Positive Definiteness** 10 credits

Let $\mathbf{A}, \mathbf{B}$ be symmetric positive definite matrices. [1] Prove that for any $p, q > 0$ that $p\mathbf{A} + q\mathbf{B}$ is also symmetric and positive definite.

Exercise 4 **General Linear Regression with Regularisation** (10+10+10+10+10 credits)

Let $\mathbf{A} \in \mathbb{R}^{N \times N}, \mathbf{B} \in \mathbb{R}^{D \times D}$ be symmetric, positive definite matrices. From the lectures, we can use symmetric positive definite matrices to define a corresponding inner product, as shown below. From the previous question, we can also define a norm using the inner products.

$$\begin{aligned} \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} &:= \mathbf{x}^T \mathbf{A} \mathbf{y} \\ ||\mathbf{x}||_{\mathbf{A}}^2 &:= \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} \\ \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{B}} &:= \mathbf{x}^T \mathbf{B} \mathbf{y} \\ ||\mathbf{x}||_{\mathbf{B}}^2 &:= \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{B}} \end{aligned}$$

Suppose we are performing linear regression, with a training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, where for each i, $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$. We can define the matrix

$$\boldsymbol{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times D}$$

and the vector

$$\mathbf{y} = [y_1, \ldots, y_N]^T \in \mathbb{R}^N.$$

We would like to find $\boldsymbol{\theta} \in \mathbb{R}^D, \mathbf{c} \in \mathbb{R}^N$ such that $\mathbf{y} \approx \mathbf{X}\boldsymbol{\theta} + \mathbf{c}$, where the error is measured using $\| \cdot \|_{\mathbf{A}}$. We avoid overfitting by adding a weighted regularization term, measured using $||\cdot||_{\mathbf{B}}$. We define the loss function with regularizer:

$$\mathcal{L}_{\mathbf{A},\mathbf{B},\mathbf{y},\boldsymbol{X}}(\boldsymbol{\theta}, \mathbf{c}) = ||\mathbf{y} - \boldsymbol{X}\boldsymbol{\theta} - \mathbf{c}||_{\mathbf{A}}^2 + ||\boldsymbol{\theta}||_{\mathbf{B}}^2 + \|\mathbf{c}\|_{\mathbf{A}}^2$$

For the sake of brevity we write $\mathcal{L}(\boldsymbol{\theta}, \mathbf{c})$ for $\mathcal{L}_{\mathbf{A},\mathbf{B},\mathbf{y},\boldsymbol{X}}(\boldsymbol{\theta}, \mathbf{c})$.

For this question:

[1] A matrix is *symmetric positive definite* if it is both symmetric and positive definite.

- You may use (without proof) the property that a symmetric positive definite matrix is invertible.
- We assume that there are sufficiently many non-redundant data points for $\mathbf{X}$ to be full rank. In particular, you may assume that the null space of $\mathbf{X}$ is trivial (that is, the only solution to $\mathbf{Xz} = \mathbf{0}$ is the trivial solution, $\mathbf{z} = \mathbf{0}$.)

1. Find the gradient $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathbf{c})$.
2. Let $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathbf{c}) = \mathbf{0}$, and solve for $\boldsymbol{\theta}$. If you need to invert a matrix to solve for $\boldsymbol{\theta}$, you should prove the inverse exists.
3. Find the gradient $\nabla_{\mathbf{c}} \mathcal{L}(\boldsymbol{\theta}, \mathbf{c})$.

   We now compute the gradient with respect to $\mathbf{c}$.
4. Let $\nabla_{\mathbf{c}} \mathcal{L}(\boldsymbol{\theta}) = \mathbf{0}$, and solve for $\mathbf{c}$. If you need to invert a matrix to solve for $\mathbf{c}$, you should prove the inverse exists.
5. Show that if we set $\mathbf{A} = \mathbf{I}, \mathbf{c} = \mathbf{0}, \mathbf{B} = \lambda\mathbf{I}$, where $\lambda \in \mathbb{R}$, your answer for 4.2 agrees with the analytic solution for the standard least squares regression problem with L2 regularization, given by

$$\boldsymbol{\theta} = (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^T \mathbf{y}.$$